**Hastings Law Journal**

Volume 59 | Issue 5 Article 7

1-2008

# Communicating Opinion Evidence in the Forensic Identification Sciences: Accuracy and Impact

Dawn McQuiston-Surrett

Michael J. Saks

# Communicating Opinion Evidence in the Forensic Identification Sciences: Accuracy and Impact

DAWN MCQUISTON-SURRETT* AND MICHAEL J. SAKS**

## INTRODUCTION

Forensic identification evidence is presented in criminal trials in various ways. Experts commonly testify that markings from a crime scene "match," "are consistent with," or "are similar to" a known person or object, with the implication being that the defendant is the source of the evidence. How forensic identification experts express their observations, how they express their opinion, how they explain what it means, and what they say it implies, can be expected to have important effects on what fact finders conclude from the evidence. The aim of this Article is to consider how the import of the results of forensic identification examinations can be most accurately and effectively communicated to trial fact finders.

At least three problems typically confront the communication of forensic identification findings. First, obstacles make reaching correct results difficult for any given examination. With the principal exception of DNA typing, virtually all areas of forensic identification lack empirically and statistically meaningful measures of the probability that questioned crime-scene marks and known suspect exemplars share a common origin. Examiners are, at present, unable to compute random match probabilities; instead, they assume that the pool of candidates in the population, which can match as well or better than the known suspect, equals precisely one. So, if they find two markings to be indistinguishably alike, they assume that they "share a common origin"

* Assistant Professor of Psychology, Division of Social and Behavioral Sciences Division, Arizona State University; Ph.D.,, University of Texas at El Paso, 2003.

** Professor of Law, Professor of Psychology, and Faculty Fellow, Center for the Study of Law, Science, & Technology, Arizona State University; Ph.D., Ohio State University, 1975; M.S.L., Yale Law School, 1985. We wish to thank Jonathan Koehler for his contribution to this research. Special thanks are extended to the Maricopa County, Arizona Superior Court for their generous assistance to us in the data collection process, and to Mitchell Bartholomew, Rebecca Bautista, Allyson Horgan, Kristina Papp, and Danielle Pieters for their help with data collection. This research was supported by a grant from the Project on Scientific Knowledge and Public Policy.

"to the exclusion of all others in the world" and that they have therefore "identified the source."[1] That such a conclusion is scientifically impossible does not prevent it from being the dominant paradigm of twentieth century forensic identification.[2]

Second is the challenge of communicating the results of forensic identification examination accurately, without error, exaggeration or intentionally or unintentionally misleading fact finders. For example, some experts reporting the results of DNA typing, rather than reporting the random match probability (RMP) or a likelihood ratio, or some other indication of the probability associated with the finding that the suspect DNA and the crime scene DNA shared certain attributes, instead state that they have "identified" the suspect as being the source of the crime scene DNA.[3] For another example consider the field of microscopic hair identification, the one field of traditional forensic identification that acknowledges its inability to pinpoint the source of questioned hair.[4] In their courtroom testimony, however, some hair examiners present testimony which exaggerates the ability of the technique to zero in on a source.[5] These examples indicate inaccurate portrayals of the identifying power of the examination and its results.

Third, is the problem of reporting accurate results in a way that enables fact finders to appreciate the meaning of the results and which enables them to incorporate the forensic identification information with other identification-relevant trial evidence so that the likelihood is maximized that correct ultimate conclusions about identity are reached.[6]

---

1. *See, e.g.*, United States v. Green, 405 F. Supp. 2d 104, 116 (D. Mass. 2005).

2. *See generally* DAVID J. BALDING, WEIGHT-OF-EVIDENCE FOR FORENSIC DNA PROFILES (2005); Christophe Champod & Ian W. Evett, *A Probabilistic Approach to Fingerprint Evidence*, 51 J. FORENSIC IDENTIFICATION 101 (2001); Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCIENCE 892 (2005); Michael J. Saks & Jonathan J. Koehler, *The Individualization Fallacy in Forensic Science*, 61 VAND. L. REV. 199 (2008) [hereinafter Saks & Koehler, *Individualization Fallacy*]; William C. Thompson & Simon A. Cole, *Psychological Aspects of Forensic Identification Evidence*, *in* EXPERT PSYCHOLOGICAL TESTIMONY FOR THE COURTS 31 (Mark Costanzo et al. eds., 2006).

3. An RMP is the probability that a person or object selected at random from the population would have the same attributes as the crime scene person or object; a likelihood ratio is the ratio of the probability of a match if the DNA in the evidence sample and that from the suspect came from the same person to the probability of a match if they came from different persons. *See* FEDERAL JUDICIAL CENTER, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 534, 573 (2d ed. 2004).

4. *See* ANDRE MOENSSENS ET AL., SCIENTIFIC EVIDENCE IN CIVIL AND CRIMINAL CASES (4th ed. 1995).

5. *See* John I. Thornton & Joseph L. Peterson, *The General Assumptions and Rationale of Forensic Identification*, *in* 4 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY § 29:37 (David L. Faigman et al. eds., 2007–2008 ed.).

6. From the perspective of pure impact, all indications are that forensic identification expert witnesses are doing very well. It has been suggested that they have been shaped over the years to offer testimony in terms that are highly influential with fact finders. *See* Michael J. Saks, *Merlin and Solomon: Lessons from the Law's Formative Encounters with Forensic Identification Science*, 49 HASTINGS L.J. 1069, 1080–94 (1998). But what the law expects of witnesses is to convey full

This Article discusses how fact finders interpret and respond to the expert testimony of forensic science examiners. Part I is an analysis of the words to be used when forensic experts report their findings to fact finders. Parts I and II describe empirical studies we have done which examine variations in the way that forensic expert testimony has been or could be presented, in an effort to explore how their testimony can be made most informative. In Part III, with the help of research exploring other areas of communication, as well as forensic communication, we try to discover ways in which the communication of forensic identification examination results might be improved. In Part IV, we review the relevant literature on fact finders' interpretation of statistical and probability evidence as it applies to forensic identification, and the extent to which opposing experts and cross-examination counter the influence of an expert's testimony. The overarching goal of the Article is to explore the communication of forensic identification science in the courtroom in order to try to ensure that fact finders can make the best and most accurate use of such evidence.

## I. THE MEANING OF WORDS

Various fields of forensic science have begun to promulgate standards to guide their member practitioners in conducting examinations, reaching conclusions, and communicating conclusions to fact finders. Some of these standards involve the words to be used when reporting findings to fact finders.

For example, the American Board of Forensic Odontology (ABFO) has developed a set of terms they expect their members to use, along with definitions of what those terms mean, so that forensic dentists will know when to use which term in characterizing the conclusions of their examinations.[7] Table 1 presents those terms and their definitions. But do those words convey the same meaning to the audience of lay fact finders (be that a judge or a jury) that they have to the expert putting them forward? It is important to the integrity of judicial decisions to ensure that fact finders understand the terminology used by experts in the same way that the experts intended them.

To find out, we asked 183 undergraduate students to indicate the meaning of the terms used by forensic odontologists. They were asked to indicate on a 100-point scale what they took the expert's intended meaning to be in regard to how certain it was that crime scene evidence originated from the suspect.[8]

---

information honestly, accurately, and usefully, so that fact finders can reach the most nearly accurate conclusion on those facts and in their verdicts.

7. *See generally* AM. BD. OF FORENSIC ODONTOLOGY, BITEMARK STANDARDS AND GUIDELINES (2006).

8. In relevant part, the questionnaire stated: The list of terms you will be encountering in this

TABLE I: TESTIMONIAL TERMS AS DEFINED BY THE AMERICAN BOARD OF FORENSIC ODONTOLOGY AND AS UNDERSTOOD BY LAYPEOPLE

| TESTIMONY | OFFICIAL DEFINITION | JURY INTERPRETATION |
|---|---|---|
| Reasonable Scientific Certainty | Highest order of certainty; no reasonable probability of error | 70.7 |
| Probable | More likely than not; most people could not leave such a mark | 57.4 |
| Consistent (with) | Similarity, but no degree of specificity, like match; may or may not be | 75.6 |
| Match | Some concordance, some similarity, but no expression of specificity intended; generally similar but true for large percentage of population | 86.0 |

The averages of the responses to each of the terms is given in the third column of Table I. We can see that the term that forensic odontologists adopted to indicate the strongest connection between source and crime scene evidence, "reasonable scientific certainty," scored a rating of 70.7. Respondents thought this term was a moderately strong expression of a common source, but it was exceeded by two other terms. "Consistent with" (a term that was adopted to mean quite a weak linkage between crime scene evidence and the suspect) is defined as having "similarity, but no degree of specificity . . . may or may not" share a common source. This term was interpreted by laypersons as indicating a stronger connection (75.6) between evidence and source than "reasonable scientific certainty." And respondents hear the term "match," intended to indicate the weakest linkage ("no expression of specificity intended; generally similar but true for large percentage of population") as indicating the *strongest* association (86.0) between crime scene evidence and its source.[9] Finally, "probable," which was intended to be the term indicating the second strongest association, was interpreted as reflecting the lowest association (57.4). In all, the terms indicated to laypersons approximately the opposite of what the experts intended the terms to mean.

study are used by forensic scientists to express to a jury the degree of confidence they have that a sample of something taken from a suspect and a sample left at the crime scene by the perpetrator came from one and the same person. In the questionnaire that follows, we will be asking you to indicate how much confidence it seems to you is being expressed by each of the different terms that forensic scientists use.

9. The rules direct odontologists to define "match" when they use it. *See* AM. BD. OF FORENSIC ODONTOLOGY, *supra* note 7. Whether many or any in fact follow this guidelines, and whether jurors successfully change their understanding of the word "match" in response to being told the formal definition, is unknown.

These findings suggest a straightforward lesson. Forensic expert witnesses cannot simply adopt a term, define for themselves what they wish it to mean, and expect judges and juries to understand what they mean by it. Of course this is better than the days, not long ago, when each expert said whatever he or she wished to say, with no formal guidelines from the community of fellow examiners. But more is needed if the courts are to be assured that fact finders will understand the terms to mean what they are intended to mean. What is needed is empirical testing of the responses to the words. As shown by the study above, such empirical testing need not be difficult.

The study just described tests individual words or phrases, presented outside of the context of case facts. Studies of greater richness, context, and complexity can test reactions to more complete versions of expert testimony, as described in the next section.

## II. JUDGES' AND JURORS' INTERPRETATIONS OF FORENSIC IDENTIFICATION TESTIMONY

Since forensic identification—outside of DNA typing[10]—has no random match data to share with fact finders, such experts generally assert that their examination indicates that the crime scene evidence and the defendant's sample "share a common source," or some comparable formulation.[11] The research described in this section is concerned with the effects of the kind of testimony that is given for the numerous other forensic identification sciences that have been in existence much longer than DNA typing has—among them fingerprints, handwriting, microscopic hair comparison, bitemarks, tire marks, shoe prints, firearms comparison, and so on. These fields have no scientific basis of the sort that DNA typing does, and must stand instead on the foundation of the personal experience and judgment of their practitioners, who must support their claims in court not with empirical data but with testimonial assurances. In order to determine the effects of more complete and candid descriptions of how experts reached their findings, we conducted two experiments.

In two studies, we tested the effects of different ways of communicating the same forensic science expert evidence to judges and to jurors.[12] These included: (1) asserting that the crime scene hair and the defendant's hair were a "match"; (2) that they were "similar-in-all-

10. DNA typing refers to the methods used to analyze DNA sequences for purposes of identification.

11. A number of examples are provided in Saks & Koehler, *Individualization Fallacy*, *supra* note 2.

12. *See generally* Dawn McQuiston-Surrett & Michael J. Saks, The Testimony of Forensic Identification Science: What Expert Witnesses Say and What Factfinders Hear (2008) (unpublished manuscript, on file with authors).

microscopic-characteristics"; (3) that they were indistinguishably alike and the examiner made a subjective estimate of the probability that the defendant was the source; (4) that they were indistinguishably alike and the examiner had the data to make an objective estimate of the RMP and expressed that in a single-probability frame; (5) or that they were indistinguishably alike and the examiner had data to make an objective estimate of the RMP and expressed that in a multi-frequency frame. In addition, we studied two other phenomena: (1) the possible effects of an examiner giving an opinion on the ultimate issue of identity; and (2) informing jurors about limitations of the expertise. All of the above can be considered as falling along a continuum ranging from more fully and candidly informing fact finders about the nature and process of forensic identification to giving them little more than conclusory assertions. We measured fact finders' inferences about the likelihood the defendant was the source of the crime scene evidence and fact finders' assessment of their own understanding of the forensic testimony.

We chose microscopic hair identification as the vehicle for exploring the effects on fact finders' judgments of various ways that forensic examiners' findings could be presented. Do fact finders draw a different understanding from the conventional qualitative conclusion ("the hairs are similar-in-all-microscopic-characteristics") versus the examiner's use of the more culturally weighted term "match"? What happens if instead of a conventional qualitative conclusion the examiner were to make explicit the probabilistic nature of the inferences that can be drawn from the examination by offering the probability of a random match? Does it make a difference if the RMP is a subjective guesstimate or an objective calculation based on empirical evidence? We were interested in what happens when a field moves from no RMPs at all to giving RMPs. Does it matter whether those RMPs are subjective guesses rather than objective calculations derived from actual data?

### A. Study One

Our first study examined the impact on judges and jurors of variations in the presentation of the forensic expert's findings. First, we varied the language and concepts by which the expert communicated the results of his examination. This study involved two groups of participants: judges and jurors.[13] Participants were presented with a case summary of a murder trial, the focus of which was the testimony of a microscopic hair examination expert who asserted in one way or another that the defendant's hair and the crime scene hair matched.[14]

13. Of the 425 participants, 128 were judges and other judicial officials attending a statewide judicial conference in Arizona, and 297 were venirepersons (prospective jurors from which a jury is selected) called for jury duty in Maricopa County Superior Court.

14. The case involved the following details: A convenience store clerk was murdered during the

In different conditions of our experiment, the expert on microscopic hair examination presented his findings in one of five different ways: (1) he characterized the similarity between the crime scene evidence and the defendant's hair as being a *match*; (2) he characterized the findings as showing the hair samples to be *similar-in-all-microscopic-characteristics*; (3) he gave his *subjective probability* estimate of the RMP; (4) he gave an objective probability estimate framed as *single-probability*; or (5) he gave an objective probability estimate framed as *multiple-frequency*.[15]

In half of the conditions described above, the forensic expert's testimony went no further. In the other half, the expert went on to offer an ultimate conclusion on identity: "The examiner concluded by offering the opinion that, based on his examination of the hair in this case, that the defendant . . . was the source" of the sample. Once the examiner properly shares his underlying findings with the court, the fact finders know as much as the expert does about the contribution of the hair evidence to the case; fact finders can connect it to the rest of the evidence in the case and reach a conclusion about identity and guilt. After reading over the case summary, participants answered questions concerning the case.

Results showed that jurors estimated a higher average probability that the defendant was the source of the hair (68%) than did judges (48%). Also, participants inferred a higher probability that the defendant was the source of the crime scene hair when the expert testimony was presented in the form of "match" (66%), "similar-in-all-microscopic-characteristics" (68%), or as an objective single-probability (69%), than when it was presented in a subjective-probability (41%) or objective multiple-frequency format (45%). When asked how much the hair evidence contributed to the defendant's guilt (on a seven-point scale where 1 = not at all, and 7 = extremely much), judges' mean responses did not vary as a function of presentation type, whereas jurors found the expert's findings to be more persuasive when they were presented as an objective single-probability (5.37) than as a subjective probability (4.42)

course of a robbery. No surveillance video was available. A bystander witnessed the crime but did not get a good look at the perpetrator's face. A murder weapon was not recovered. Police identified several potential suspects who agreed to provide a hair sample for comparison with the hair recovered from a ski mask dropped at the crime scene. This analysis linked one of the suspects to the crime scene. Thus, the critical evidence in the case was the comparison of the perpetrator's hair from the crime scene with hair sampled from the defendant.

15. In the *subjective probability* version, the expert offered a subjective guesstimate concerning the rarity of the hair left at the crime scene, expressed quantitatively. In the objective *single-probability* version, the expert expressed the rarity of the hair left at the crime scene in purely quantitative (probability) terms framed in such a way to focus on the defendant. In the objective *multiple-frequency* version, the expert expressed the rarity of the hair left at the crime scene in purely quantitative (frequency) terms framed in such a way to focus on the larger population.

or an objective multiple-frequency (4.51).[16]

We also asked our judges and jurors how many people would have hair that is indistinguishably similar to the hair recovered from the crime scene in a city of 500,000 people, in light of the expert testimony. The correct answer to the question is 500, which was readily accessible to participants presented with quantitative testimony.[17] Nearly half of the participants in these conditions gave exactly the correct answer of 500: 41% of those in the subjective probability condition and 46% of those in the objective multi-frequency condition. In the objective single-probability version of the testimony in which the arithmetic was slightly harder, 25% of participants gave the correct answer.[18] Those three conditions resulted in estimates that were a fraction of the size of the estimates given by participants in the remaining two conditions—"match" and "similar in all microscopic" characteristics—in which guessing was unavoidable since no quantitative information was provided.

It is interesting to note that participants in the conditions which led to the highest estimates that the crime scene hair came from the defendant paradoxically gave the highest estimates of the incidence of the same hair traits in the reference population. This reinforces the inference that those two testimonial conditions lead to the least understanding of the basic concepts of forensic identification while leading to the highest inculpatory judgments.

While jurors' appraisals of their own understanding of the expert's testimony (on a seven-point scale) did not vary as a function of presentation format, judges on average felt that they better understood the expert evidence when it was presented as an objective multiple-frequency (4.73) than in the "similar in all microscopic" characteristics condition (3.16). In general, the highest ratings of understanding tended to result from the more quantitative presentations.

Whether or not the expert gave an opinion on the ultimate issue of identity had no impact on any of the responses of the participants to the rest of the evidence.

### B. Study Two

Based on the results of our first study, our second study had two goals. First, it sought to discover whether the understanding and impact

16. Wherever we state that a difference was found, a test of statistical significance showed a difference at a probability level of ($p < 0.05$).

17. Under the objective multi-frequency and subjective probability conditions, the expert stated that the number would be 3,000 in a city of 3,000,000, so arriving at the proportionate number in a city of 500,000 was a matter of simple arithmetic.

18. The expert stated that the incidence rate in the population was 0.001, so the participants would need to multiply 500,000 by 0.001.

of forensic identification testimony could be improved by providing information to the jurors about the limitations of such expertise. Specifically, jurors learned of the scientific limitations of forensic identification either through cross-examination by the defense attorney or by an instruction from the judge. Second, the previous experiment had failed to find any effect of the forensic expert offering an ultimate opinion or not, so we sought to make the expert's rendering of an ultimate opinion more salient. In this study we looked at only three forms of presentation: match and similar-in-all-microscopic-characteristics are the two that are most commonly encountered under current practice, and subjective probability is the only remotely possible option for the near future.[19]

Participants consisted of 350 venirepersons called for jury duty from the same county court as in the previous experiment. Participants were presented with the same basic murder trial used in the first study. One independent variable involved informing the jury about limitations of microscopic hair examination. In a control condition, jurors were not informed about any limitations. In another version, limitations were brought out on cross-examination.[20] In a third version the limitations were presented by the judge.[21]

For the experiment's second independent variable, the expert either gave no ultimate opinion or the case summary stated: "Asked what bottom-line conclusion his examination of the hair led to, the forensic expert stated that the hair found in the ski mask most likely belonged to the defendant, and therefore it was Aaron Robinson who had been wearing that ski mask at the robbery."

The third independent variable involved three different forms in

19. Calculations of objective RMPs or likelihood ratios must wait until the day when sufficient population data on forensically relevant attributes are gathered so that they can be used in routine casework.

20. The following was included in the case summary:

> The defense attorney then cross-examined the forensic hair examiner. The attorney asked whether the expert's opinion could be taken to reflect any particular degree of accuracy, and the witness answered that it could not. The attorney asked whether the assumptions underlying the expert's opinion had been subjected to thorough scientific testing, and the witness answered that there had been little scientific testing. The attorney asked whether it was not true that the expert's opinion amounted to little more than his subjective judgment, and the expert answered that his conclusions were his subjective judgment informed by his experience working on previous cases. The attorney asked whether the present case would become part of that "experience working on previous cases," and the witness acknowledged that it would.

21. The following was included in the case summary:

> As part of instructing the jury, the judge cautioned that, by its nature, the expert testimony presented in this case lacks any particular degree of accuracy because it has never been tested scientifically. The conclusions of the expert are little more than his subjective judgment. The court noted that it nevertheless thought that the testimony had enough value to be considered by the jury. The court concluded by reminding the jurors that they alone have the authority and the responsibility to give the expert testimony as much or as little weight as they feel it deserves.

which the forensic expert reported his findings: match, similar-in-all-microscopic-characteristics, and subjective probability (identical to what was presented in the first study). After reading the case summary participants answered similar questions as participants in the first study.

Similar to our earlier findings, results from our second study indicate that participants were most persuaded by the match (74%) and the similar-in-all-microscopic-characteristics versions of the testimony (70%) in evaluating the probability that the crime-scene hair came from the defendant. Inferences that the defendant was the source of the crime-scene hair were also higher when the expert offered an explicit conclusion that the defendant was the source (55%) than when he did not offer this conclusion (34%), but only in the subjective-probability condition. Why might this be? Perhaps testimony in the form of match and similar-in-all-microscopic-characteristics already elevates jurors' estimates of source probability, so for them the explicit conclusion has little to add. But jurors hearing the subjective-probability explanation, thereby made aware of the guesswork involved in the expert's opinion, are less sure of what to conclude (reflected in their lower estimates of source probability). For jurors in this condition the expert's ultimate-opinion statement bolsters his testimony, making up for the less persuasive (because it is a more complete and accurate portrayal of the nature of the expertise) subjective-probability testimony.

Two additional questions probed the impact of the expert testimony on inferences that the defendant was in fact the perpetrator. When we asked the jurors how sure they were that the defendant committed the crime, on a seven-point scale, jurors in the match (4.50) and similar-in-all-microscopic-characteristics (4.32) conditions were on average more sure than those in the subjective probability condition (3.69). Paralleling the findings described above, the effect of giving or not giving an explicit conclusion interacted with the form of testimony such that ultimate-opinion testimony increased belief in the defendant's guilt, but only when the testimony had been given in the subjective-probability form. We asked jurors about the contribution of the expert evidence to the strength of the case against the defendant, also on a seven-point scale. Again, ultimate-opinion testimony made a difference only when subjective probabilities were given: jurors thought the hair evidence was stronger when a conclusion was given (4.83) than when it was not given (3.94). The main lesson of these findings is the unshakeableness of the traditional forms: match and similar-in-all-microscopic-characteristics produce something of a ceiling effect, which resist moderation by the presentation of other information.

Jurors rated their understanding of the expert testimony as greater on average when an explicit opinion on the ultimate issue of identity was given (5.26) than when no ultimate issue testimony was given (4.90).

Offering an ultimate opinion on the possible implications of the examination cannot logically improve a juror's understanding of how the examination was conducted, but it apparently led jurors to feel that they better understood it. Most likely, by hearing the expert assert an ultimate conclusion, the jurors felt less uncertainty about the implications of the expert's findings, and they transferred that feeling of comfort with a "verdict" to their sense of their actual understanding of the testimony and its basis.

Jurors' understanding of the expert evidence as reflected in their answer to the question of whether a forensic scientist's "conclusion that the suspect is the person who left the evidence at the crime scene would be strongest when the size of the pool of others who would match is" smaller or larger did not vary as a function of any of our independent variable manipulations. But it is noteworthy that responses to this question were decidedly in the correct direction: an overall mean of 2.95 on a seven-point scale. By this measure, then, jurors generally seemed to grasp one of the major concepts underlying forensic identification.

Whether or not jurors were informed about the limitations of microscopic hair examination on cross-examination or by the judge had little measurable or meaningful impact on their judgments about the likelihood that the defendant was the source of the crime-scene hair or their perceived understanding of the expert's testimony.

### C. Implications of Our Findings

Overall, our research generally found that jurors were more influenced by the expert's testimony than were judges, arriving at higher probability estimates that the defendant was the source of the crime scene evidence, and being somewhat more influenced than the judges were by the form of the expert testimony presented. In examining an objective measure of understanding of a key concept of forensic identification (calculating the occurrence rate in the population of crime scene evidence with a given set of attributes), the jurors were not less accurate than the judges except in the single-probability condition (which required more calculations); in that condition 48% of judges reached exactly correct answers compared to only 17% of jurors. In the two conditions where judges and jurors could only guess at the population rate, judges made guesstimates that were far higher than those of the jurors (which would be consistent with the differences in their respective estimates of source probability). Neither judges nor jurors were influenced by whether the expert asserted an explicit opinion on the ultimate issue of the identity of the defendant as the source of the crime scene evidence.

We also found evidence that the two traditional forms[22] of testimony—"match" and "similar-in-all-microscopic-characteristics"—generally behaved in tandem. They both lead to high source probability estimates and to high estimates of the population rate of the crime scene evidence. This finding that these two forms of testimonial expression had similar effects is noteworthy because "match" is considered by many microscopic hair examiners to be a misleading characterization of hair comparison findings, and "similar-in-all-microscopic-characteristics" to be the more enlightened substitute. These traditional forms in which forensic identification testimony is expressed do not seem to differ in their impact on jurors or judges, presumably communicating a comfortingly simple and easily grasped (though not very informative and presumably misleading) understanding of the basis for the identification opinion. While these traditional forms of forensic identification testimony produced the highest estimates of source probability, the most plausible and candid alternative, namely, the subjective-probability form of testimony, produces the lowest estimates.

Since most jurors have an exaggerated view of the nature and capabilities of forensic identification, we expected that information explaining limitations of the expertise would temper the jurors' inferences; but the information on limitations had little meaningful effect on jurors' judgments. We also expected that when an expert gives an explicit ultimate opinion that a defendant was the source of crime scene evidence, fact finders would be more persuaded that the defendant was the source than when the testimony did not include that ultimate opinion. We found some evidence for this in our second study: explicit conclusions by the expert increased source-probability estimates and certainty that the defendant was the perpetrator, but only when the expert testimony was of the subjective-probability type. The traditional forms of testimony may be so robust as to create something of a ceiling effect which renders other testimonial elements, such as an explicit ultimate opinion, largely superfluous. But where the more modest subjective-probability testimony is presented, room remains to boost the fact finders' inferences, and then the assertion of an explicit conclusion makes a difference. Giving an ultimate opinion on identity also increased jurors' assessments of their own understanding of the expert testimony. Such responses, however, are the reason that the common law ultimate opinion doctrine prohibited experts from opining on ultimate issues: it

22. These traditional terms have been used to express expert opinion on identification (often sounding more like fact) for many decades, and continue in use. *See generally* Saks & Koehler, *Individualization Fallacy*, *supra* note 2. Other familiar and common phrases include: "share a common source," "identification," and "to the exclusion of all others in the world." *Id.* It should be said that "match" is not regarded as appropriate by many or most microscopic hair identification experts, but it is used by other identification fields, which is why we tested it in these experiments. *Id.*

tended to invade areas of decision reserved for fact finders. But modern rules of evidence have abolished the ultimate issue rule, except regarding certain psychological issues in criminal cases, where the prohibition lives on.[23] Perhaps trials would benefit from expanding what remains of the rule against opining on ultimate issues.

### D. Our Additional Research on the Expression of Forensic Analyses

We conducted additional research that tested whether the bare conclusion of a forensic expert was all that matters to fact finders, or whether knowing something additional about his background or the methods he used to conduct his analyses made a difference to the meaning or acceptability of his conclusions concerning the evidence.[24] This study examined whether knowing the simplicity and subjectivity of the examination process enhances or vitiates fact finders' belief in the conclusion the expert witness draws.

Participants[25] read a case that paralleled the one used in our experiments described earlier, but it was presented in the form of a written transcript of the trial instead of a summary. The details of the case involved a convenience store clerk who was murdered in the course of a robbery. The transcript revealed that an individual made an eyewitness identification under poor conditions which led to the arrest and prosecution of a man who had been found in the neighborhood of the crime and loosely fit the description. Further, the defendant's brother testified that they were at his home around the time of the robbery. The critical evidence is the comparison of hair taken from a ski mask dropped at the crime scene by the perpetrator with hair sampled from the defendant. Thus, the focus was on the testimony of a microscopic hair examination expert who asserted a link between the defendant's hair and the crime scene.

The study's design consisted of variations in the expert's years of experience in the field, and variations in the description he gave of the forensic hair examination process.[26] In each of these the expert gave an ultimate conclusion as to the identity of the perpetrator based on the results of his analysis. We also included a control condition which removed the above four variations wherein the expert presented only a conclusion without offering any information about his background or the

23. *See* FED. R. EVID. 704.

24. The study discussed in this and the next three paragraphs describes an honors thesis conducted under the supervision of the authors of the present Article. *See* Bianca Connolly, The Effects of Forensic Science Expert Testimony: How Little Is Enough? (2006) (unpublished honors thesis, Arizona State University) (on file with authors).

25. Participants included 152 Arizona State University undergraduate students.

26. The expert expressed having either one or twenty-three years of experience, and then he either provided a brief description of the hair examination process or he did not.

process of his examination. It seemed possible that jurors would be prepared to trust the findings of the more experienced expert, especially when they learned that the process depended entirely on the subjective judgment of the expert. But it was also possible that jurors would treat all versions of the testimony equally because the expert's bottom line was the most meaningful piece of his assertions.

When we asked participants to evaluate the expert, the assertion of years of experience of that expert had little impact. However, jurors who heard him describe the hair examination process (especially compared to those who heard only the bare conclusion with no mention of the process or the expert's credentials) rated the expert as doing a significantly better job, being more scientific, being more convincing, and facilitating a better understanding of his analyses. But their judgments on the actual evidence (i.e., likelihood/probability/certainty that the hair came from the defendant, helpfulness of the evidence in the case against the defendant) and verdict did not vary as a function of hearing the examination process information. Thus, while differences between conditions occurred for jurors' beliefs that they were more impressed by the expert who explained the hair examination process to them, the data do not suggest that they were actually more influenced by that expert. And what the jurors learned about the hair examination process should not have been very comforting to them: it should have revealed how little science there was behind his analysis and that his conclusion was based on nothing more than an intuitive, subjective guesstimate of whether the hairs share a common source.

These studies examined the extent to which fact finders' judgments are affected by the various ways a forensic examiner can describe his findings. The next section explores the parallels that can be drawn between the field of forensic identification and other fields which seek effective communication of important information to decision makers.

## III. Research on Communicating Risk

If forensic identification testimony can be viewed as testimony which conveys information about a "risk"—that an accused might not be the perpetrator, that a real perpetrator might still be on the loose, or that the perpetrator is the defendant who might escape liability if his identity is not recognized—then research on risk communication from other fields might provide relevant insights for the presentation of forensic identification evidence.

Clinical psychologists and other mental health professionals are increasingly called upon to offer expert testimony to the courts in which they estimate the future dangerousness of a violent offender. The clinical assessment of violence risk and the degree of accuracy of such conclusions are well-researched topics in the psychological literature, but

how efficacious the communication of that risk is to others has received less attention.[27] A similar line of inquiry lies within the fields of medicine and health communication regarding how best to present patients with risk information relevant to various diseases and treatment options.[28] Within the context of the legal system there is an obvious need for the assessment of future risk in various domains to be communicated to the courts in a way that is clear, concise, complete, and usable; the assessor's evaluation must be "fully accessible" to those who are then responsible for making decisions in the courtroom.[29] The research—and much of its findings—on risk communication drawn from these fields lends itself to some of the work we have done in examining expert testimony proffered in the forensic identification sciences and how the content and results of forensic analyses can best be expressed to fact finders.

Risk communication can be thought of as the connection between the assessment of risk and what decisions should be made based on that assessment. One article appropriately argued that even the most accurate of assessments of violence risk will be useless to decision makers if its conclusions are not communicated effectively.[30] How can risk-relevant information be delivered in such a way that maximizes decision makers' understanding? This is an important question for any field in which predictions are made and consequences based on the perceived level of risk follow.

For instance, in evaluating the possible future dangerousness of a violent offender, decisions based on risk assessments can involve determining appropriate intervention strategies, civil commitment or sentencing, possible harm to others,[31] and so on. In the healthcare field, patients make decisions about screening tests, treatments, the effectiveness of certain treatments, and so on, based on their perception of the associated risks and benefits explained to them by healthcare professionals. In our field of interest—communicating the results of forensic examinations—decisions involve determining the extent of a link between crime scene material and samples taken from an individual or object, ultimately leading to an assessment of culpability.

Numerical or statistical assessments are often used to convey a level of future risk, expressed generally using a probability format or a

27. *See generally* John Monahan, *Violence Prediction: The Past Twenty Years and the Next Twenty Years*, 23 CRIM. JUST. & BEHAV. 107 (1996).

28. *See, e.g.*, Angela Fagerlin et al., *Making Numbers Matter: Present and Future Research in Risk Communication*, 31 AM. J. HEALTH BEHAV. S47, S47 (2007).

29. Robert F. Schopp, *Communicating Risk Assessments: Accuracy, Efficacy, and Responsibility*, 51 AM. PSYCHOLOGIST 939, 939 (1996).

30. Kirk Heilbrun et al., *Violence Risk Communication: Implications for Research, Policy, and Practice*, 1 HEALTH RISK & SOC'Y 91, 103 (1999).

31. That is, a possible duty to warn third parties about a patient's dangerousness. *See* Tarasoff v. Regents of the Univ. of Cal., 551 P.2d 334, 351 (1976).

frequency format. For example, a clinician might express the risk of a person committing future violence as "Person X has a 30% chance of hurting someone in the future" (a probability) or as "Of 100 persons like X, we might expect thirty to hurt someone in the future" (a frequency); these conclusions are numerically identical. Unfortunately, ample evidence from the research literature suggests that most people have poor quantitative and statistical reasoning skills both in general and, specifically, in legal settings.[32] Concerning the use of probabilities and frequencies to convey an examination's results or an assessment, studies consistently find that decision makers reach different judgments when they are presented with information in a frequency format versus the same information communicated in probability terms. Specifically, the presentation of quantitative information as frequencies leads people to estimate greater likelihoods than if they receive information in the form of a probability.[33]

In the communication of violence risk, the data show that framing a risk assessment as a frequency leads decision makers to view the person of interest as being at a greater risk of harming others than when that risk is expressed as probability.[34] For example, Slovic et al. presented clinicians with case summaries describing at-risk patients and asked them to judge the likelihood of future harm to others within six months after hospitalization discharge.[35] A major finding consistent across their three studies was that patients were judged by psychologists and psychiatrists as posing a greater risk to others when the likelihood of committing a harmful act was framed as a frequency rather than as a probability, regardless of the type of response scale used.[36] Even a tutorial given to participants that was designed to assist them in their interpretation of the quantitative information had no impact on the consistency of their judgments and, specifically, did not reduce the apparent biasing effects of the frequency versus the probability formats.[37] Monahan et al. report similar findings concerning the effect of frequency versus probability formats among a group of forensic psychologists.[38]

These effects might be due to the use of an "affect heuristic" in which the interpretation of a frequentistic assessment ("Of 100 persons

32. *See, e.g.*, Jonathan J. Koehler, *When Are People Persuaded by DNA Match Statistics?*, 25 Law & Hum. Behav. 493, 494 (2001).

33. Paul Slovic et al., *Violence Risk Assessment and Risk Communication: The Effects of Using Actual Cases, Providing Instruction, and Employing Probability Versus Frequency Formats*, 24 Law & Hum. Behav. 271, 289 (2000).

34. *Id.* at 272–73.

35. *Id.* at 285–88.

36. *Id.* at 289–90.

37. *Id.* at 289–92.

38. John Monahan et al., *Communicating Violence Risk: Frequency Formats, Vivid Outcomes, and Forensic Settings*, 1 Int'l J. Forensic Mental Health 121, 126 (2002).

like X, we might expect thirty to hurt someone in the future") may evoke frightening images of violent mental patients, leading the person being assessed to be viewed as a high risk to society, versus a less threatening image of one patient based on a probabilistic assessment ("Person X has a 30% chance of hurting someone in the future").[39] Koehler argues that this explanation can be thought of as a special case of "exemplar cueing," in that the tendency to generate exemplars of others who have the potential to commit violence results from the presentation of frequencies, leading to a more conservative risk management approach compared with that of a probabilistic estimate.[40]

In Koehler's research on the evaluation of DNA match statistics, exemplar cueing theory applies nicely to the frequency versus probability presentation effects he observed.[41] He presented laypeople with case summaries that varied the presentation format of the likelihood of a DNA match as either a frequency (e.g., the likelihood of a match if the suspect is not the source is one in 1,000) or a probability (e.g., the likelihood of a match if the suspect is not the source is 0.1%). Across several studies he found that the frequency format was far less persuasive than the probability format, resulting in lower-rated guilt judgments.[42] These results suggest that the frequentistic presentation led people to estimate a greater likelihood of a DNA match *only by coincidence*, resulting in lower ratings of guilt, thus paralleling the literature described above that finds a greater perceived likelihood of risk based on frequency assessments.[43] Results of our research examining these format effects based on the presentation of other types of forensic evidence are consistent with these findings.[44]

As an alternative to the presentation of a risk assessment in purely quantitative terms, some suggest the use of categorical risk communication that would use a classification format to convey the likelihood of future risk, based on the number and severity of risk factors present, and the like.[45] For example, in a study examining methods of risk communication, psychologists and psychiatrists preferred the use of categories to express risk ("low" vs. "moderate" vs. "high" risk of

39. *See generally* Ali Siddiq Alhakami & Paul Slovic, *A Psychological Study of the Inverse Relationship Between Perceived Benefit*, 25 RISK ANALYSIS 1085 (1994); Melissa L. Finucane et al., *The Affect Heuristic in Judgments of Risk and Benefits*, 13 J. BEHAV. DECISION MAKING 1 (1999); Paul Slovic et al., *supra* note 33.

40. Koehler, *supra* note 32, at 495–98.

41. *Id.*

42. *Id.*

43. Slovic et al., *supra* note 33.

44. *See generally* McQuiston-Surrett & Saks, *supra* note 12.

45. Kirk Heilbrun et al., *Expert Approaches to Communicating Violence Risk*, 24 LAW & HUM. BEHAV. 137, 140–41 (2000).

violence) over the use of numerical probabilities.[46] Other research is consistent with this finding, demonstrating clinicians' preference against the use of probabilities to express an assessment of future risk.[47] Arguably, an integration of numerical/probabilistic plus categorical risk assessments could in fact provide the most comprehensive set of information to aid decision makers.

A related line of inquiry has explored laypersons' perceptions of mental health practitioners' expert testimony when predictions of risk are based on clinical opinion (subjective judgments) versus actuarial instruments (objective judgments).[48] Opinion testimony is that which is based on the clinician's experience in practice, whereas prediction originating from an actuarial assessment is based on scientifically established risk factors shown to predict violence among groups of offenders (e.g., prior record of violence, criminal history, psychopathic assessment).[49] Although assessments based on actuarial methods are often more accurate at predicting future violence than are clinical judgments,[50] the research described below finds that laypeople in fact are more persuaded by clinical opinion expert testimony than actuarial expert testimony when it comes to predicting dangerousness.

In a study examining the judgments of mock jurors, Krauss and Sales presented videotaped simulated testimony that manipulated the type of expert testimony presented (clinical vs. actuarial), along with various adversarial procedures (cross-examination, competing experts) designed to safeguard biasing testimony.[51] Their results showed that participants were more influenced by clinical opinion testimony.[52] Mock jurors were more persuaded in their ratings of future dangerousness by the clinical testimony compared with the actuarial testimony both prior to and following implementation of any adversarial procedure, and found the two forms of testimony to be equally scientific and credible.[53]

Some parallels between these findings and our own research on fact finders' evaluation of forensic science expert testimony can be drawn. In our research, when we asked participants to indicate the likelihood of a forensic match between crime scene material and that taken from a

46. *Id.*

47. *See generally* Kirk Heilbrun et al., *Risk Communication: Clinicians' Reported Approaches and Perceived Values*, 27 J. AM. ACAD. PSYCHIATRY L. 397 (1999); Charles Lidz et al., *The Accuracy of Predictions of Violence to Others*, 269 JAMA 1007 (1993).

48. Daniel A. Krauss & Bruce D. Sales, *The Effects of Clinical and Scientific Expert Testimony on Juror Decision-Making in Capital Sentencing*, 7 PSYCHOL. PUB. POL'Y & L. 267, 269 (2001).

49. *Id.* at 271–73.

50. Randy Borum, *Improving the Clinical Practice of Violence Risk Assessment*, 51 AM. PSYCHOLOGIST 945, 950–52 (1996).

51. Krauss & Sales, *supra* note 48, at 282–85.

52. *See id.* at 289–99.

53. *Id.*

defendant, participants were often more influenced by the forensic expert's subjective (clinical), qualitative (nonquantified) judgment describing the methods he used to arrive at his conclusion of a match, compared to a more objective and quantitative description of his analysis.[54]

More generally, the findings on risk communication from the health and mental health fields likely find their parallels in forensic science expert testimony. The conclusions of examiners in all areas of forensic identification other than DNA typing reach their conclusions on the basis of subjective guesstimations (clinical rather than actuarial), they present their opinions in nonquantitative, usually categorical, terms, and by all indications laypersons are generally quite persuaded by their testimony.[55]

We next turn to a broad discussion of fact finders' decision making concerning expert witnesses. Our review describes research examining the impact of experts' conclusions which are often framed in quantitative terms on fact finders' reasoning, the expression of forensic analyses and conclusions as framework evidence, and whether the legal safeguards designed to counter the influence of experts' testimony have a real impact in the courtroom.

## IV. Expert Witnesses and Decision Making by Fact Finders

A considerable amount of research on jurors' (and sometimes judges') understanding of and reactions to expert testimony has been conducted in contexts other than the conventional, low-technology forensic science on which this paper has been focused. In this section we review what we regard as some of the most informative of that literature for the purposes of our topic.

### A. Statistical and Probability Evidence

One implication of our own research[56] is that trial fact finding might be better served if forensic scientists would testify more fully and could provide an empirical, perhaps even quantitative, basis for their testimony. After all, that is how so much valuable scientific data are often conveyed to the makers of important decisions. But as suggested by some of our research, fact finders are not always sensitive to such differences in the testimony presented.[57]

Relatedly, several lines of research suggest that statistical and

54. *See generally* McQuiston-Surrett & Saks, *supra* note 12.

55. *See, e.g.*, State v. Quintana, 103 P.3d 168, 171 (Utah Ct. App. 2004) (Thorne, J., concurring) ("In essence, we have adopted a cultural assumption that a government representative's assertion that a defendant's fingerprint was found at a crime scene is an infallible fact, and not merely the examiner's opinion.").

56. *See generally* McQuiston-Surrett & Saks, *supra* note 12.

57. *Id.*

probability evidence present special difficulties for fact finders.[58] In general, probability evidence tends to be underutilized by fact finders and they tend not to be sensitive to variations in some important parameters,[59] though the pattern of findings is not simple. Changes in features of the trial process and in the format by which statistical evidence is presented produce changes in fact finders' understanding and use of the evidence.

One of the most relevant sets of studies is by Thompson and his colleagues on the "Prosecutor's Fallacy" and the "Defense Attorney's Fallacy."[60] These can occur when an expert opines that a match exists between crime scene evidence and evidence known to originate with the defendant, and data are offered so that the fact finders have some basis for evaluating the likelihood that the two samples shared a common source, namely, the defendant. As an example, a long blond hair is found at a crime scene and a suspect with long blond hair is arrested. The significance of the long blond hair depends, to an important degree, on how common or how rare that trait is in the population; such evidence would be more probative in China than it would be in Sweden.

In the Prosecutor's Fallacy, fact finders mistakenly think that the frequency of the trait in the population tells them something about the probability of guilt or innocence of the particular defendant.[61] For example, if fact finders learn that a trait shared by the perpetrator and the defendant occur in 2% of the population, many of them infer that the chance that the defendant is not the source is only 2%. In the Defense

58. *See generally* David H. Kaye & Jonathan J. Koehler, *Can Jurors Understand Probabilistic Evidence?*, 154 J. ROYAL STAT. SOC'Y 75 (1991); William C. Thompson, *Are Juries Competent to Evaluate Statistical Evidence?*, 52 LAW & CONTEMP. PROBS. 9 (1989).

59. *See* Michael J. Saks & Robert F. Kidd, *Human Information Processing and Adjudication: Trial by Heuristics*, 15 LAW & SOC'Y REV. 124 (1981); *infra* notes 60–69. William C. Thompson and Simon A. Cole raise an important criticism of some of the jury simulation findings:

> [S]ubjects' apparent conservatism in these early studies may have been due, in part, to the inadequacy of the Bayesian models. In these models the likelihood ratio depended on a single variable—the random match probability. Although these likelihood ratios may have reflected the diagnostic value of the forensic match, they failed to capture any uncertainty about its reliability. In other words, the Bayesian models against which subjects' judgments were compared implicitly assumed the forensic tests were error-free. This is a big assumption and one that subjects probably did not share.

Thompson & Cole, *supra* note 2, at 53–54. Later and better studies find less underutilization (greater "accuracy"), but fact finders still fall short of the Bayesian norm. Bear in mind, also, that in circumstances where laypersons underutilize probability data, so does everyone else, including statisticians and scientists. *See generally* DANIEL KAHNEMAN ET AL., JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES (1982). The implication of much of this kind of research (outside of jury studies) is that even professional decision makers are prone to reaching incorrect conclusions if they rely on their intuition rather than making use of explicit computational decision aids.

60. *See* Thompson, *supra* note 58; William C. Thompson & Edward Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167 (1987).

61. Forensic science experts themselves sometimes make this error. *See* Thompson, *supra* note 58, at 26–27.

Attorney's Fallacy, fact finders might realize that a population rate of 2% would mean that in a city of, say, 1,000,000 people, 20,000 would have the trait, and mistakenly regard the evidence as having virtually no probative value on the question of identity. The first error overvalues the evidence and the second undervalues it.[62]

Other studies have found that when mock jurors are given the same statistical evidence just described, but in the context of conditional probabilities (mock jurors were told that if the defendant were innocent there was a 2% chance that his trait would match the perpetrator's), more of them committed the Prosecutor's Fallacy and fewer committed the Defense Attorney's Fallacy (8%).[63] When the trait frequencies were given as a percentage and an incidence rate (2% of people have hair that would be indistinguishable, and in a city of 1,000,000 there would be approximately 20,000 such individuals), then fewer jurors committed the Prosecutor's Fallacy but more committed the Defense Attorney's Fallacy.[64]

Other research varied the frequency of the trait in the population (5%, 1%, 0.1% or no probability information at all).[65] The research found that jurors' estimates "failed to make fine distinctions between probability estimates that were mildly incriminating, moderately incriminating, and strongly incriminating."[66] The mean estimate of guilt, however, was higher in groups that heard the probabilistic evidence than in the control group that heard no data, implying that when no data are given or can be given, fact finders implicitly substitute their own guesstimates of the likelihoods.[67]

In a study in which counsel on both sides made invalid arguments on behalf of the two fallacies, most jurors thought that either the prosecutor's fallacy (29%) or the defense attorney's fallacy (68%) was correct, while only 22% of the jurors concluded that both arguments were incorrect. Moreover, jurors' erroneous decisions suggest that "it is easy to talk people into using inappropriate judgmental strategies to

62. Eight percent of jurors or fewer make these errors. *See* Jane Goodman, *Jurors' Comprehension and Assessment of Probabilistic Evidence*, 16 AM. J. TRIAL ADVOC. 361, 384 (1992); Thompson & Schumann, *supra* note 60, at 174.

63. Thompson & Schumann, *supra* note 60, at 174.

64. *Id.*

65. Goodman, *supra* note 62, at 371.

66. *Id.* at 371–72. In the 5% condition, the mean estimate of guilt was 40%, in the 1% condition it was 45%, and in the 0.1% condition it was 47%. *Id.* at 371. Goodman compares the student answers to the answers provided by the Bayes' Theorem. Using the conviction rate of the control group (where no probabilistic evidence was presented) as the prior estimate of guilt, compared to a Bayesian rational juror, the mock jurors in each group tended to underutilize the blood type evidence. *Id.* at 372. The discrepancy was greatest in the groups in which the frequency probabilities were the most incriminating. *Id.* at 373.

67. *Id.* at 371.

evaluate [this kind of] evidence."[68] When jurors were permitted to discuss the evidence and arguments, as they would in deliberation, the estimates of probability of guilt and their conviction rates dropped. The exception to this were those exposed to the fallacious arguments for the prosecutor's fallacy, but this was the case only when the defense did not counter that with an argument in favor of the defense attorney's fallacy.

Typically, fact finders give insufficient weight to population rate data that have been offered to clarify the meaning of a "match."[69] They do increase their estimates of the probability of guilt in response to the evidence, but not as much as Bayes' Theorem suggests their guilt judgments ought to be adjusted.[70]

Research by Koehler has shown, on the other hand, that how exactly one presents mathematically equivalent evidence can have quite dramatic effects on the inferences jurors draw from the data.[71] Mentioned above, and replicated in our own experiments, Koehler's experiments tested the effects of varying the target and the frame of DNA match statistics.[72] When describing the chance of a coincidental match to an innocent person, the focus (the target) of the presentation can be on the defendant in the courtroom or, say, the metropolitan population. And the incidence rate of the trait in the population can be expressed ("framed") as RMP or as frequency.[73] To take the two most different combinations: An example of a multi-target frequency frame would be to inform fact finders that, although the defendant's DNA matches the crime scene DNA, this would be true also for one in one million people in the city where the crime occurred.[74] An example of a single-target probability frame would be to inform fact finders that the chance that the suspect would match by coincidence if he were not the source is 0.000001.[75] As Koehler notes, these two kinds of reports are mathematically identical, but psychologically different.[76] Upon hearing the evidence by way of a single-target probability frame, fact finders are far more likely to infer that the defendant was the source of the DNA than when the same information is communicated by way of a multi-target frequency frame.[77] Differences as a function of the incidence rate

68. Thompson, *supra* note 58, at 33.

69. Koehler, *supra* note 32, at 493.

70. *Id.* at 504; Thompson, *supra* note 58, at 33.

71. Koehler, *supra* note 32, at 509.

72. *Id.* at 497–98.

73. *Id.* at 497.

74. *Id.*

75. *Id.*

76. *Id.*

77. *Id.* at 498. In one of the experiments, where the incidence rate in the population was one in 1,000, 54% of mock jurors hearing the single-target probability frame thought the defendant was more than 99% likely to be the source of the crime scene DNA, while the same evidence presented through

in the population (one in a thousand versus one in a million versus one in a billion), however, had only modest impact on jurors' judgments.[78]

When jurors are made aware of the laboratory's error rate in determining whether a questioned and a known were indistinguishably alike, or were thought to share a common source, how do fact finders combine such information with the incidence rate of the trait in the population and what effect do both together have on their estimations of the probability of identification? For example, a juror who initially estimates the probability of the defendant's guilt to be only 10%, and who then receives new evidence indicating that the defendant and perpetrator have the same blood type, which occurs in 1% of the population, and that the false-positive error rate of matching blood types is also 1%, should, according to the Bayesian model, arrive at a new combined probability of guilt of about 85%.[79] If the incidence rate and false-positive rate are both 5%, then the Bayesian model indicates the revised estimate of probability of guilt should be only 53%. The difference between 1% and 5% may seem small at first blush, but the impact on the probative value of the match between the defendant and perpetrator is substantial. But to reach the proper result, a fact finder must evaluate two probabilities at once: population incidence rate and laboratory error rate.

In the first of Thompson's experiments testing fact finders' ability to evaluate the evidence described above, mock jurors were given several different sets of evidence to examine, so that they could compare cases with stronger and weaker evidence.[80] They accurately determined that the evidence was most incriminating where both probabilities were low, least incriminating where both probabilities were high, and of intermediate incriminating value where one factor was high and the other low. In the next experiment, experimenters gave participants only one case to evaluate.[81] Presented in this way, which is more like the way jurors actually encounter cases, fact finders were insensitive to the differences in the probativeness of the evidence.[82] In a third experiment they were again presented with the evidence of one case but now were allowed to deliberate in groups (like juries) about the evidence.[83] With deliberation, jurors were still insensitive to the differences in the evidence—by finding the evidence quite probative whether it was or was

---

a multi-target frequency frame led 36% of jurors to conclude that there was less than a 1% chance that the defendant was the source of the DNA. *Id.* at 506.

78. *Id.* at 508.

79. Thompson, *supra* note 58, at 36.

80. *Id.* at 30–31. This is what researchers call a within-subjects design.

81. *Id.* at 34. This is what researchers call a between-subjects design. Now the comparisons are between groups of jurors receiving one kind of evidence compared to another.

82. *Id.*

83. *Id.* at 37.

not. In the strong evidence condition jurors reached estimates of guilt (66%) that came very close to the Bayesian calculation (64%), while jurors in the weak evidence condition greatly overestimated (65%) the real value (22%) of the weak evidence.[84]

A way out of the dilemma of either withholding from jurors meaningful data or giving them data with which they might lead themselves astray is to find ways to present such evidence in a manner that will facilitate correct use. Nance & Morris, using a large number of jurors drawn from an Illinois jury pool, compared jurors who were presented with entirely nonquantitative DNA match findings in the context of a criminal case, to jurors who were given quantitative data in one of several different forms to see which best aided them in assessing the probativeness of the DNA match.[85] Jurors were given data on the frequency in the population of the DNA sequence at issue and data on the frequency of false positive laboratory errors.[86] The population frequency data were given in the form of a frequency (one in 40,000), or as a frequency plus as a likelihood ratio (40,000 times more likely to match if the accused is the source of the crime scene sample than if he were not), or as a frequency plus a likelihood ratio plus a chart that mapped how such a likelihood ratio should change people's judgments of the probability of guilt depending on how incriminating they thought the other evidence in the case was.[87] The evidence on laboratory error was presented either in unquantified form ("there is a chance of lab error"); or quantified lab error ("one false positive in every thousand tests"); or given, in addition, with instruction on the proper way to combine laboratory error with the population frequency evidence. Jurors made best use of the random match population data when they were provided with the frequency data, the likelihood ratio, and the chart; they made the least effective use of the evidence when presented with only the frequency data. Information about laboratory error had no appreciable impact on fact finders' judgments.[88]

Although underutilization errors are made in regard to unemotional issues in calm contexts, emotional arousal probably contributes to the problem of undervaluation of probabilistic evidence. Sunstein refers to this as "probability neglect."[89] When strong emotions are involved,

84. *Id.*

85. *See* Dale A. Nance & Scott B. Morris, *Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random-Match Probability*, 34 J. LEGAL STUD. 395, 401–04 (2005).

86. *Id.* at 402–03.

87. *Id.*

88. *Id.* at 409–10.

89. *See* Cass R. Sunstein, *What's Available? Social Influences and Behavioral Economics*, 97 Nw. U. L. REV. 1295, 1303 (2003); Cass R. Sunstein, *Probability Neglect: Emotions, Worst Cases, and Law*, 112 YALE L.J. 61, 61–76 (2002).

people pay less attention to information and apparently even less attention to probability.[90] Large scale differences in probability that should matter seem to matter little at all in an individual's decision.

## B. FRAMEWORK EVIDENCE

Forensic identification evidence is a specie of "framework" evidence. Framework evidence consists of background information on persons or events not parties to the case at bar, but who arguably are similar to, or in a situation similar to, that of the relevant party.[91] Studies of framework evidence have typically involved such issues as eyewitness unreliability, posttraumatic stress disorders, or cross-cultural differences in the meaning of behavior.[92] The purpose of the testimony is to provide the fact finder with general information about the context in which contested adjudicative facts occurred in order to help the fact finder interpret the contested case-specific facts.[93]

Thus, forensic identification evidence informs (or should inform) a fact finder of the relationship between characteristics of objects in the relevant universe of objects which are not at issue on the present case, which would then inform them about the attributes of the evidence at issue in the present case, and it would be up to the jury to apply the general knowledge to the particulars at issue in the case. If that were done when forensic identification evidence were presented, and if the existing research in those rather different areas were applicable, the research on "social framework" expert testimony would suggest that fact finders give considerable weight to such evidence.[94]

In some kinds of cases, illustrated by the topics in the studies cited in the preceding footnote, experts disclaim any ability to opine on the

---

90. For a general discussion of the role of emotion in jury decision making, see Jeremy A. Blumenthal, *Law and the Emotions: The Problems of Affective Forecasting*, 80 IND. L.J. 155 (2005), and Reid Hastie, *Emotions in Jurors' Decisions*, 66 BROOK. L. REV. 991 (2001).

91. Laurens Walker & John Monahan, *Social Frameworks: A New Use of Social Science in Law*, 73 VA. L. REV. 559, 572–73 (1987).

92. *Id.* (explicating cases and studies involving these and other examples of such "framework" uses of evidence).

93. Though less obvious than in cases where framework evidence is proffered, these are the same kinds of facts contained in all sorts of scientific studies on which all kinds of practitioners rely: one understands the specific (this patient, this bridge, this vehicle, this bee hive, this forest, etc.) mostly through one's understanding of many others of the same category that have been studied earlier. In other trial contexts, this background knowledge is what makes the expert an expert, and it is a major (often unspoken) foundation on which the expert's opinion stands.

94. *See generally* Natalie J. Gabora et al., *The Effects of Complainant Age and Expert Psychological Testimony in a Simulated Child Sexual Abuse Trial*, 17 LAW & HUM. BEHAV. 103 (1993); Regina A. Schuller & Sara Rzepa, *Expert Testimony Pertaining to Battered Woman Syndrome: Its Impact on Jurors' Decisions*, 26 LAW & HUM. BEHAV. 655 (2002); Neil Vidmar & Shari Seidman Diamond, *Juries and Expert Evidence*, 66 BROOK. L. REV. 1121 (2001); Neil Vidmar & Regina A. Schuller, *Juries and Expert Evidence: Social Framework Testimony*, 52 LAW & CONTEMP. PROBS. 133 (1989).

implications of the general knowledge to the case at bar. They offer the scientific findings, and let the fact finders apply them to the specifics of the case.[95] Other experts are willing to tell the fact finders what conclusion the general knowledge leads to in the specific case being litigated. The law has been inconsistent concerning when it will allow which kind of expert testimony.[96] The abolition of the bar against ultimate issue opinions in most situations would seem to have opened the door to the case-specific conclusions, which the research shows will have a greater impact on the jury. Experts on forensic identification usually give an ultimate opinion on identity, rather than to offer the jury framework evidence plus "clinical" information on the features of the evidence involved in the instant case. Thus, framework phenomena do not have an opportunity to play themselves out in testimony by forensic identification scientists.

### C. WELLS' "BLUE BUS" STUDY

Gary Wells set out to test a number of explanations for the famous Blue Bus Problem. The usual finding, in research subjects as well as judicial opinions, is that people hesitate to make liability decisions when the only available evidence consists of naked statistics.[97] Legal scholars tend to debate the legal justification for and the philosophical underpinnings of such hesitancy. Wells attacked the problem as a matter of cognition, empirically testable.

The basic scenario is that a plaintiff, who is color blind, is suing the Blue Bus Company for killing her dog. The sort of evidence she has to offer is naked statistics: county transportation officials testify that there are only two bus companies in the county, the Blue Bus Company and the Grey Bus Company, and that the Blue Bus Company owns 80% of all the buses and generates 80% of all bus traffic on the road. Faced with such evidence, most people (judges, college students, MBAs, and others) make approximately correct subjective probability estimates of the likelihood that it was a Blue Bus Company bus that ran over the dog, but only a fraction of them are willing to find the Blue Bus Company liable. By contrast, a witness who testifies that he thinks he saw the Blue Bus do it, even though the witness is correct only 80% of the time, is capable of leading judges and students to find liability.[98]

The following scenarios are most relevant to our line of research.

---

95. *See* Walker & Monahan, *supra* note 91, at 559.

96. *Id.*

97. Gary L. Wells, *Naked Statistical Evidence of Liability: Is Subjective Probability Enough?*, 62 J. PERSONALITY & SOC. PSYCHOL. 739, 739 (1992). Wells uses the term "naked statistical evidence" to mean "probabilities that are not case specific in the sense that the evidence was not created by the event in question but rather existed prior to or independently of the particular case being tried." *Id.*

98. *Id.* at 744.

One body of evidence consisted of tire tread patterns from the accident scene that are then compared to the buses in the Blue and Gray company fleets. An expert explains that 80% of the buses that had tire treads that matched the pattern left at the accident scene were owned by the Blue Bus Company while 20% were owned by the Gray Bus Company. In a slightly different version, the expert adds that, based on the tire tread evidence, "he believed that the bus that ran over Mrs. Prob's dog was a Blue Bus Company bus." With this one small addition, judges as well as students were willing to find against the defendant. Judges and students receiving both versions reach the same subjective probability estimates of the likelihood that the Blue Bus Company is the offender. But verdicts against the Blue Bus Company were rendered only in the presence of the latter testimony.

These findings would seem to suggest that an important reason that jurors so often accept the liability implications of forensic identification testimony is that examiners are permitted to make conclusory assertions based on their own subjective judgment. According to Wells's theory, by permitting witnesses to give case-specific, fact-to-evidence (rather than requiring evidence-to-fact testimony), findings of liability become far more common.[99]

### D. Opposing Experts and Cross-Examination

The Krauss and Sales research, discussed above, is one of a few studies that directly assesses the ability of opposing experts and cross-examination to counter the influence of an expert's testimony.

Diamond and her colleagues examined this proposition within the context of a criminal case.[100] The stimulus involved the testimony of an expert modeled after a Dr. James Grigson, a psychiatrist who regularly testified for the prosecution in death penalty cases on the issue of future dangerousness.[101] Typically, Dr. Grigson would testify that the defendant constituted an ongoing danger. For example, in *Barefoot v. Estelle* he stated that there was a "one-hundred percent and absolute chance" the defendant would commit future crimes of violence.[102] In the Diamond et al. experiment, the jury, drawn from the Cook County, Illinois jury pool, watched a seventy-five minute videotape of a death penalty hearing involving an armed robbery and murder of a stranger whom the

99. *Id.* at 746.

100. Shari Seidman Diamond et al., *Juror Reactions to Attorneys at Trial*, 87 J. Crim. L. & Criminology 17, 35 (1996).

101. *Id.* at 36. In Texas, where Dr. Grigson most frequently testified, the jury could not impose the death sentence unless they concluded that the defendant was likely to "commit criminal acts of violence that would constitute a continuing threat to society." Tex. Code Crim. Proc. Ann. art. 37.071(2)(b)(1) (Vernon 2002).

102. Barefoot v. Estelle, 463 U.S. 880, 919 (1983).

defendant robbed in order to buy beer.[103]

In three conditions of the experiment, the jurors heard the prosecution expert state that he had diagnosed the defendant as a sociopath, based solely on an examination of records of prior court proceedings, pre-sentence reports, and prison records.[104] The expert concluded that the defendant was "certain to kill again" if not executed. The expert asserted that he had extensive prior experience in making such predictions and he was generally accurate.

In the first, "weak cross-examination" condition, the defense only brought out the fact that the expert usually testified for the state, but did not challenge the future dangerousness prediction.[105] In the second, "strong cross-examination" condition, the defense added to the cross in the weak condition by pointing out at length that the expert's prediction of future killing was inconsistent with prior research and that the expert has not employed the standard methods for diagnosing future dangerousness. The expert admitted on the stand that the best scientific literature indicates that two-thirds of dangerousness predictions prove to be incorrect. The cross-examination also brought out the fact that the expert had never published his findings in peer-reviewed journals. The expert responded that he was focused on clinical diagnosis, not publication, and that he was confident he was correct. In the third, "strong cross-examination plus defense expert" condition, the defense lawyer conducted the same cross as in the strong-cross condition. In addition, a defense expert, who was also a psychiatrist, testified that the defendant coped reasonably well but on rare occasions excessive drinking interacted with a personality disorder to produce violence. The defense expert testified that predictions about future violence could not be made with any certainty, but that in his view the likelihood of future similar violence was not great and the defendant was a good candidate for an alcohol abuse program.

In a fourth, "control" condition, the prosecution expert made a realistic prediction, basically agreeing with the defense expert that predictions of future dangerousness are accurate only about one-third of the time, but warned about the defendant's potential for future violence. The cross-examination was identical to the cross in the first condition.

The dependent variables[106] in the study included a question about the persuasiveness of the state's expert, the jury verdict preference (death or life) and a verdict confidence index. The first condition, with a

103. Diamond et al., *supra* note 100, at 19–20. The next three paragraphs summarize the method section of the article.

104. *Id.* at 38.

105. *Id.* It is not unusual for lawyers to devote a large portion of their cross-examination of experts to the issues of the expert's qualifications and potential sources of bias.

106. *Id.* at 38–42.

strong prediction of future dangerousness, no opposing experts and a weak cross should produce the highest percentage of death penalty verdicts. If cross-examination is an effective prophylactic against unreliable testimony, the second condition should produce lower persuasiveness scores and a lower percentage of death penalty verdicts. And the combination of a powerful cross and an opposing expert should produce still lower persuasiveness scores and even fewer death penalty verdicts. Ideally, this version would produce jury judgments indistinguishable from the fourth version in which the expert reported a one in three chance of being correct.

In fact neither the "strong cross" nor the "strong cross plus the opposing expert" had a significant effect on plaintiff expert persuasiveness, percentage of juries opting for death, or verdict confidence.[107] For example, in the weak cross condition 47% of the juries gave a death verdict, in the strong cross condition 51% recommended death, and in the strong cross plus opposing expert 50% recommended death. The only condition with a different result was in the fourth, "control" condition where the plaintiff expert testified that predictions were wrong two-thirds of the time. In this condition, 39% of the juries recommended the death penalty.[108] Diamond and Casper note that one possible interpretation of these results is that the jurors simply did not care about future dangerousness.[109] However, based on evidence from their deliberations, this is not the case. Most juries explicitly discussed the issue and there was a strong correlation between jurors' predictions of future dangerousness and verdict preferences.[110] However, jury estimates of future dangerousness if released did not vary significantly across the three conditions where the plaintiff's expert testified the defendant would kill again.[111]

This one study does not establish the inefficacy of "battles of the experts" or cross-examination. Diamond and Casper offer the possibility that this testimony was particularly difficult to overcome because it was consistent with beliefs and expectations already held by the jurors.[112] However, the results are consistent with another study, by Kovera and colleagues.[113] They varied the strength of the defense's cross-examination of an expert. Although appropriate tests revealed that jurors were

107. *Id.* The results described in this paragraph summarize the "Results" section of Diamond et al.

108. *Id.* at 40.

109. *Id.*

110. *Id.* at 43.

111. *Id.* at 42.

112. *Id.* at 53. There is research supporting the proposition that mock jurors hold strong beliefs concerning the ability of clinicians to predict future dangerousness and that they overestimate clinician accuracy. *See* Daniel Krauss & Bruce Sales, *supra* note 48, at 276.

113. *See* Margaret Kovera et al., *Expert Testimony in Child Sexual Abuse Cases: Effects of Expert Evidence Type and Cross-Examination*, 18 LAW & HUM. BEHAV. 653, 653 (1994).

sensitive to the relative strength of the cross-examination of the expert, this did not affect participants' perceptions of the quality of the evidence nor did it affect verdict.[114] This result was replicated in a second study by the same authors.[115] These results should give pause to anyone who believes that the traditional tools of the adversarial process will always undo the adverse effects of weak expert testimony.

## CONCLUSION

The subjective nature of the state of the art of most fields of forensic identification is generally not well understood. Our research findings show that the traditional forms of testimony (similar-in-all-microscopic-characteristics, match, and equivalent formulations) were the most damaging to the defense, leading judges and jurors alike to high estimates of source probability. Presumably this is not a coincidence, but the result of decades of "practice" in seeing what leads to the most acquiescence by fact finders. These forms of testimony do not, however, lead to the greatest understanding of what forensic identification does and means.

One might have expected an explication of the examination process, emphasizing the guesswork involved, would have a sobering effect on fact finders, but it appears instead to lead fact finders to be more impressed by the examination. Similarly, since most jurors begin with an exaggerated view of the nature and capabilities of forensic identification, one might expect that information explicitly informing fact finders about the limitations of the expertise would temper the jurors' inferences. Such information had little effect on jurors' judgments.

Since fact finders, especially jurors, tended to yield to comforting certainties of expression about the evidence being testified to, one might also expect that when an expert gives an explicit ultimate opinion that a defendant was the source of crime scene evidence, fact finders would be more persuaded that the defendant was the source than when the testimony did not include that ultimate opinion. Ultimate opinion testimony only increased jurors' assessments of their own understanding of the expert testimony.

Both important similarities and important differences are found in the patterns of response by judges and jurors. Jurors were more influenced by the expert's testimony than judges, arriving at higher probability estimates that the defendant was the source of the crime scene evidence. Judges appear to be less affected than jurors by subtle variations in the form of presentation. Both judges and jurors are more

114. *Id.* at 669.

115. *See* Margaret Kovera et al., *Reasoning About Scientific Evidence: The Effects of Juror Gender and Evidence Quality on Juror Decisions in a Hostile Work Environment Case*, 84 J. APPLIED PSYCHOL. 362, 369–71 (1999).

comfortable converting subjective probability evidence into findings of liability when the expert asserts a personal interpretation of a conclusion to which the data point. Neither judges nor jurors were influenced by whether the expert asserted an explicit opinion on the ultimate issue of the identity of the defendant as the source of the crime scene evidence. Lay fact finders of most kinds generally have difficulty understanding statistical, and especially probability, data, and underutilize such information.

Clearly, the language employed by forensic experts affects the inferences fact finders draw, sometimes producing conclusions in the minds of fact finders quite different from what the expert witnesses purportedly intend. More subtly, even slight variations in how an expert's testimony is structured or the words used can significantly affect the understanding fact finders' draw from it. And, unfortunately, cross-examination and the use of opposing experts do not appear to effectively counter expert testimony, regardless of the logical vulnerability of the initial expert testimony.

The findings on risk communication from the health and mental health fields seem to find their parallels in the communication of forensic identification science expert testimony. The conclusions of examiners in all areas of forensic identification other than DNA typing reach their conclusions on the basis of subjective guesstimations (arguably clinical rather than actuarial), and they present their opinions in nonquantitative, qualitative, usually categorical, terms. These seem to be the very attributes of risk communication (and forensic psychological testimony) that have more influence on decision makers. The paradox is that systematic, data-based, quantitative evidence might be viewed by genuine scientists as the best way to support and express information, but those are less persuasive (to lay fact finders and perhaps to all kinds of fact finders) kinds of testimony than testimony which scientists hold in higher regard—and which all rational decision makers would, in an ideal world, would also hold in higher regard.

Much remains to be learned about how fact finders understand and respond to the expert testimony of examiners in the various forensic individualization fields, and how their testimony can be made most informative. The patterns that have emerged from the studies that have been conducted thus far both advance our understanding of fact finder decisions concerning forensic identification science and can serve to guide the path of future research.

***